Estimation of parameters in incomplete data models defined by dynamical systems.

Adeline Samson, Sophie Donnet

Estimation of parameters in incomplete data models defined by dynamical systems

Sophie Donnet[1], Adeline Samson[2]

[1] Paris-Sud University, Laboratoire de Mathématiques, Orsay, France
[2] INSERM U738, Paris, France; University Paris 7, Paris, France

Abstract

Parametric incomplete data models defined by ordinary differential equations (ODEs) are widely used in biostatistics to describe biological processes accurately. Their parameters are estimated on approximate models, whose regression functions are evaluated by a numerical integration method. Accurate and efficient estimations of these parameters are critical issues. This paper proposes parameter estimation methods involving either a stochastic approximation EM algorithm (SAEM) in the maximum likelihood estimation, or a Gibbs sampler in the Bayesian approach. Both algorithms involve the simulation of non-observed data with conditional distributions using Hastings-Metropolis (H-M) algorithms. A modified H-M algorithm, including an original Local Linearization scheme to solve the ODEs, is proposed to reduce the computational time significantly. The convergence on the approximate model of all these algorithms is proved. The errors induced by the numerical solving method on the conditional distribution, the likelihood and the posterior distribution are bounded. The Bayesian and maximum likelihood estimation methods are illustrated on a simulated pharmacokinetic nonlinear mixed-effects model defined by an ODE. Simulation results illustrate the ability of these algorithms to provide accurate estimates.

KEYWORDS: Bayesian estimation, Incomplete data model, Local linearization scheme, MCMC algorithm, Nonlinear mixed-effects model, ODE integration, SAEM algorithm

1 Introduction

When a biological or physiological process is measured, the regression function of the statistical model corresponding to the observed data is often derived from a differential equation describing the underlying dynamic process. Difficulties arise when the differential equation has no analytical solution and/or when the parameters of the regression function are random and thus non-observed. Such example can be found in pharmacokinetics, which aims to study drug evolutions in human organism, this evolution being described by differential systems of compartment interactions. Mixed models, for which regression parameters are considered as random variable and non-observed data, are widely used for the analysis of pharmacokinetic datasets, which have classically repeated measurements in several patients.

This paper aims at providing a general answer to the estimation problem in such statistical incomplete data models.

Let y be the noised observations of a biological process measured at instants $(t_1, \cdots, t_J)$. The biological process is described by the solution g of an ordinary differential equation (ODE), depending on a stochastic non-observed parameter ϕ:

$$y_j = g(t_j, \phi) + \varepsilon_j \qquad \text{for } j = 1 \cdots J.$$

We consider that the observable vector Y is part of a so-called complete vector (Y, ϕ). We assume that both Y and (Y, ϕ) have density functions, $p_Y(y; \theta)$ and $p_{Y,\phi}(y, \phi; \theta)$ respectively, depending on a parameter θ belonging to some subset Θ of the Euclidean space $\mathbb{R}^q$. The estimation of the parameter θ has been widely studied when the regression function g has an explicit form. Two approaches can be followed to tackle this challenge, respectively the maximum likelihood and the Bayesian estimations.

Generally, the maximization of the likelihood of the observations cannot be done in a closed form. Dempster et al. (1977) propose the iterative Expectation-Maximization (EM) algorithm for incomplete data problems. At the k^{th} iteration, the E-step of EM algorithm computes $Q(\theta|\theta_k) = E(\log p_Y(y; \theta)|y; \theta_k)$ while the M-step determines θ_{k+1} maximizing $Q(\theta|\theta_k)$.

For cases where the E-step has no closed form, stochastic versions of EM are introduced. Celeux and Diebolt (1985) introduce the Stochastic EM algorithm (SEM). Wei and Tanner (1990) suggest the Monte-Carlo EM (MCEM) estimating $Q(\theta|\theta_k)$ by the averaging of m Monte-Carlo replications. Recently, Wu (2004) emphasizes that MCEM is computationally intensive. As an alternative, Delyon et al. (1999) propose the Stochastic Approximation EM algorithm (SAEM) replacing the E-step by a stochastic approximation of $Q(\theta|\theta_k)$. These methods require the simulation of the non-observed data ϕ. For cases where this simulation can not be performed in a closed form, Kuhn and Lavielle (2004) suggest to resort to iterative methods such as Monte Carlo Markov Chain algorithms (MCMC).

The Bayesian approach estimates the posterior distribution $p_{\theta|Y}(\cdot|y)$ of θ, a prior $p_\theta(\cdot)$ being given. Because of the conditional independence structure of $p_{\theta|Y} = \int p_{\theta|Y,\phi}\ p_{\phi|Y}\ d\phi$ and $p_{\phi|Y} = \int p_{\phi|Y,\theta}\ p_{\theta|Y}\ d\theta$, Gelfand and Smith (1990) propose a Gibbs sampling to evaluate these two integrals simultaneously. At iteration k, ϕ_k, a realization of ϕ, is simulated with $p_{\phi|Y}(\cdot, \theta_{k-1})$ followed by θ_k, a realization of θ with $p_{\theta|Y,\phi}(\cdot, \phi_k)$. Consequently, as in maximum likelihood estimation, difficulties arise when the simulation of the conditional distribution can not be performed in a closed form. For these cases, a Hastings-Metropolis (H-M) algorithm can be included in the Gibbs sampler.

The use of the H-M algorithm in estimation algorithms requires the evaluation of the regression function g at each iteration. When g is a non-analytical solution of a dynamical system, it is evaluated using a numerical integration method. Thus a trade-off between accuracy, stability and computational cost is required. In this paper, we detail the Local Linearization scheme (see e.g. Biscay et al., 1996; Ramos and García-López, 1997; Jimenez, 2002) not only because of its stability performances but also because this scheme can be extended to a so-called modified Local Linearization scheme, adapted to its inclusion in the H-M algorithm. The estimation algorithms are then applied to an approximate model whose regression function is an approximate solution of the ODE.

The objective of this research is to quantify the error induced by the numerical approximation of the regression function g. The paper is organized as follows. Section 2 defines the original statistical model; the Local Linearization scheme and its modified version are detailed; the approximate statistical model resulting from the numerical approximation is introduced. Section 3 focuses on the H-M algorithm to simulate the non-observed data ϕ with the conditional distribution. The error induced by the numerical

approximation of g on the conditional distribution is quantified. Section 4 is dedicated to the parameter estimation algorithms. Concerning maximum likelihood and Bayesian estimations, the standard algorithms are adapted to solve the approximate model. The error induced by the use of the numerical solving method is bounded respectively on the likelihood and the posterior distribution. This error is distinct from the error on the estimates induced by the estimation algorithm which is evaluated by their standard errors. Finally, the SAEM algorithm and the Bayesian Gibbs sampler are applied on a nonlinear mixed-effects model deriving from pharmacokinetics in Section 5.

2 Models and notations

2.1 An incomplete data model defined by ODEs

Let $y = (y_j)_{j=1..J}$ denote the observations measured at times $(t_1, \cdots, t_J)$. We consider the incomplete data model, called model $\mathcal{M}$, defined as follows:

$$\begin{aligned} y_j &= g(t_j, \phi) + \varepsilon_j \qquad 1 \leq j \leq J \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma^2) \\ \phi &\sim \pi(\cdot; \beta) \end{aligned} \tag{$\mathcal{M}$}$$

where $g(.)$ is a nonlinear function of ϕ, ε_j represents the error of the measurement j, σ^2 is the residual variance, ϕ is a non-observed random parameter distributed with the density $\pi(\cdot, \beta)$, depending only on the parameter β. The parameter $\theta = (\beta, \sigma^2)$ belongs to some open subset $\Theta \subset \mathbb{R}^q$.

Let g be written $g = H \circ f$, where $H : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a known function and, $f : \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ is defined as the solution of the following ODE:

$$\begin{aligned} \frac{\partial f(t, \phi)}{\partial t} &= F(f(t, \phi), t, \phi) \\ f(t_0, \phi) &= f_0(\phi) \end{aligned} \tag{1}$$

with a known function $F : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ and the initial condition $f_0(\phi) \in \mathbb{R}^d$, $t \in [t_0, T]$.

We make the following additional assumptions:

- Assumption **H1**: π has a compact support $K_1 \subset \mathbb{R}^k$, and there exist two constants a and, b such that

$$0 < a < \pi(\phi; \beta) < b \text{ for all } \phi \in K_1.$$

- Assumption **H2**: $F : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^k \longrightarrow \mathbb{R}^d$ is $\mathcal{C}^2$ on its definition domain, and $\phi \longrightarrow f_0(\phi) \in \mathbb{R}^k$ is $\mathcal{C}^1$ on K_1.
- Assumption **H3**: H is an L_H-lipschitzian function.

2.2 Approximation of the regression function

A great variety of numerical schemes have been proposed to solve ODEs (see e.g. Hairer et al., 1987). The accuracy of such numerical methods is qualified by the order and the step size of these schemes. The numerical scheme is applied on sub-intervals $[t_n, t_{n+1}[$, $n = 0, \ldots, N-1$, of the time interval $[t_0, T]$, with $t_N = T$. The maximal length or step size of the sub-intervals is denoted h. The order of a numerical scheme is defined as follows:

Definition 1 *Let f_h be the resulting approximate function obtained by a numerical integration scheme of step size h. This scheme is of order p if there exists a constant C such that*

$$\sup_{t \in [t_0, T]} |f(t, \phi) - f_h(t, \phi)| \leq C h^p.$$

Of all the numerical schemes, the Local Linearization (LL) scheme provides a good trade-off between computational cost and numerical stability, as exposed in Biscay et al. (1996). It derives from the local linearization of the right term of the ODE (1) with respect to time t, and the exact integration of the deduced linear differential equation. Its implementation requires matrix exponential computations using new algorithms such as Pade or Schur methods, which have proved their efficiency and stability. The LL scheme has the additional advantage of preserving stability properties on stiff systems (see e.g. Ramos and García-López, 1997; Jimenez et al., 2002). In cases where the ODE depends on parameter ϕ, we extend this scheme using a Taylor expansion with respect to time t and parameter ϕ. More precisely, let the solution at ϕ_0 be computed using the LL scheme. Let ϕ be in a neighborhood of ϕ_0. On each sub-interval $[t_n, t_{n+1}[$, $n = 0, \ldots, N-1$, the solution f_{h,ϕ_0} of the following linear equation

$$\begin{aligned} \frac{\partial f(t,\phi)}{\partial t} &= F(f(t_n,\phi_0),t_n,\phi_0) + \frac{dF}{df}\left(f(t_n,\phi_0),t_n,\phi_0\right)\left(f(t,\phi) - f(t_n,\phi_0)\right) \\ &\quad + \frac{dF}{dt}\left(f(t_n,\phi_0),t_n,\phi_0\right)(t - t_n) + \frac{dF}{d\phi}\left(f(t_n,\phi_0),t_n,\phi_0\right)(\phi - \phi_0). \end{aligned}$$

is evaluated. Details of this scheme (called LL2 scheme) are given in appendix B. This LL2 scheme does not involve any additional computation of matrix exponentials: the required matrix exponential has already been computed with the LL scheme at ϕ_0. This reduces the computational time, which is a key issue in iterative processes. For instance, during H-M algorithm implementation, the ODE has to be integrated at ϕ contained in a neighborhood of ϕ_0, for which the ODE has already been solved by LL. This leads us to propose a modified version of the H-M algorithm, taking advantage of these schemes (see Section 3).

Convergence properties of the two previous numerical schemes are given in the following lemma. Let us consider the following assumption:

- Assumption **H1′**: ϕ remains in the compact set $K_1 \in \mathbb{R}^k$.

Lemma 1 *Let ϕ_0 and ϕ be in K_1. Let $f(.,\phi)$ be the exact solution of the ODE (1), $f_h(.,\phi)$ the one obtained by the LL scheme with step size h, and $f_{h,\phi_0}(.,\phi)$ the LL2 solution. Assume that **H1′** and **H2** hold. Then:*

1. *there exists a constant C independent of ϕ, such that, for any $t \in [t_0, T]$ and for any ϕ,*
$$|f(t,\phi) - f_h(t,\phi)| \leq Ch^2,$$
2. *there exist constants C_1 and C_2 such that, for any $t \in [t_0, T]$ and for any ϕ,*
$$|f(t,\phi) - f_{h,\phi_0}(t,\phi)| \leq \max(C_1 h^2, C_2 \|\phi - \phi_0\|^2_{\mathbb{R}^k}).$$

Part 1 is proved in Ramos and García-López (1997), part 2 is proved in appendix B.

2.3 An approximate incomplete data model

In practice, estimation algorithms require the numerical approximation of the regression function and are thus applied to an approximate version of the model $\mathcal{M}$. Let f_h be the approximate solution of the ODE (1), obtained by a numerical integration method of step size h and order p. Let the approximate statistical model $\mathcal{M}_h$ be defined by:

$$\begin{aligned} &y_j = g_h(t_j,\phi) + \varepsilon_j \qquad 1 \leq j \leq J \\ &\varepsilon_j \sim \mathcal{N}(0,\sigma^2) \\ &\phi \sim \pi(\cdot;\beta) \end{aligned} \qquad (\mathcal{M}_h)$$

where $g_h = H \circ f_h$. Subsequently, the different distributions of the model $\mathcal{M}_h$ are subscripted with h.

3 Simulation of non-observed data with the conditional distribution

In this paper, the Hastings-Metropolis (H-M) algorithm is combined successively with the SAEM algorithm in the maximum likelihood estimation, and the Gibbs sampler in the Bayesian estimation. The H-M algorithm updates ϕ in the target distribution $p(\phi|y;\theta)$. The computation of its acceptance probabilities requires an explicit expression of the regression function. As a consequence, this H-M algorithm can only be applied to the approximate statistical model $\mathcal{M}_h$.

The standard H-M algorithm and a modified version of the random-walk H-M algorithm including the LL2 scheme leading to computational time savings in practice, are presented in Section 3.1. Section 3.2 presents the convergence of the algorithms.

3.1 The Hastings-Metropolis algorithm

The iterative H-M algorithm implemented on model $\mathcal{M}_h$ generates a Markov chain with the target distribution $p_h(\phi|y;\theta)$ as invariant distribution and using a proposal density q.

At step $r+1$, given $\phi^{(r)}$:

- Generate a candidate ϕ^c from the proposal density $q(.|\phi^{(r)})$.
- Generate $U \sim \mathcal{U}([0,1])$. Then,

$$\phi^{(r+1)} = \begin{cases} \phi^c & \text{if} \quad U < \rho(\phi^{(r)}, \phi^c), \\ \phi^{(r)} & \text{if} \quad U > \rho(\phi^{(r)}, \phi^c), \end{cases}$$

where

$$\rho(\phi^{(r)}, \phi^c) = \min\left\{1, \frac{p_{h,\phi|Y}(\phi^c)}{p_{h,\phi|Y}(\phi^{(r)})} \frac{q(\phi^{(r)}|\phi^c)}{q(\phi^c|\phi^{(r)})}\right\}$$

is the acceptance probability.

The choice of the proposal density q is essentially arbitrary, although in practice a careful choice will help the algorithm to move quickly inside the parameter space. Two proposal densities are combined. First the prior density $q(.|\phi^{(r)}) = \pi(.;\beta)$ allows to move inside the parameters space efficiently. Second, a symmetric distribution $q(\phi^c|\phi^{(r)}) = q(\phi^{(r)}|\phi^c)$ is used resulting in the so-called random-walk H-M algorithm (see e.g. Bennet et al., 1996). In this case, $\phi^c = \phi^{(r)} + \delta$ with δ simulated from a centered symmetric distribution such that ϕ^c remains in K_1.

***Remark* 1** In practice*, the random-walk H-M algorithm can be modified to reduce the computational time using the LL2 scheme. Indeed, on the model $\mathcal{M}_h$ defined by the LL scheme, the random-walk H-M algorithm requires solving the ODE (1) at ϕ^c in a bounded neighborhood of $\phi^{(r)}$, for which the LL approximate solution has been computed.*

Let the centered symmetric proposal density verify the following property, called property (2): there exists $\eta > 0$ such that, almost surely,

$$\|\phi^{(r)} - \phi^c\|_{\mathbb{R}^k} < \eta. \tag{2}$$

The modified H-M algorithm including the LL2 scheme is outlined as followed. At step $r+1$, given $\phi^{(r)}$,

- *Generate a candidate $\phi^c = \phi^{(r)} + \delta$*
- *Accept ϕ^c with probability $\rho^{(2)}(\phi^{(r)}, \phi^c)$ where*

$$\rho^{(2)}(\phi^{(r)}, \phi^c) = \min\left\{1, \frac{p^{(2)}_{h,\phi|Y}(\phi^c)}{p_{h,\phi|Y}(\phi^{(r)})}\frac{q(\phi^{(r)}|\phi^c)}{q(\phi^c|\phi^{(r)})}\right\} = \min\left\{1, \frac{p^{(2)}_{h,Y|\phi}(\phi^c)\pi(\phi^c)}{p_{h,Y|\phi}(\phi^{(r)})\pi(\phi^{(r)})}\right\}$$

  and $p^{(2)}_{h,Y|\phi}(.)$ is evaluated using the LL2 scheme.

If the move is accepted $f_h(t, \phi^{(r+1)})$ and the $p_{h,Y|\phi}(\phi^{(r+1)})$ density are re-evaluated using the LL scheme. If it is not accepted no additional matrix exponential is computed, significantly reducing the computational time.

3.2 Convergence

The convergence of the standard H-M algorithm and the error induced by the numerical integration scheme are studied in Theorem 1.

Theorem 1 *Let f be the exact solution of ODE (1). Let $p_{\phi|Y}$ be the conditional distribution for the model $\mathcal{M}$. Assume that **H1**, **H2** and **H3** hold. Let f_h be the approximate solution obtained by a numerical integration method of step size h and order p. Let $p_{h,\phi|Y}$ be the conditional distribution of the model $\mathcal{M}_h$.*

1. *Then on the model $\mathcal{M}_h$, the H-M algorithm converges towards its stationary distribution $p_{h,\phi|Y}$.*

2. *Furthermore, there exists a constant C_y such that for any small h,*

$$D(p_{\phi|Y}, p_{h,\phi|Y}) \leq C_y h^p.$$

where $D(\cdot,\cdot)$ denotes the total variation distance.

The rates of convergence of the standard H-M algorithms have been widely studied (see e.g. Tierney, 1994), hence are not discussed here.

***Remark* 2** *The modified H-M algorithm based on the LL2 scheme results in a non standard form of the acceptance probability due to the two approximations of the target distribution. Thus, the proof of its convergence is complex and beyond the scope of this paper. In cases where the modified H-M algorithm converges to a stationary distribution $p_h^{(2)}$, provided the property (2) is checked and the chain $(\phi^{(r)})$ remains in K_1, there exists a constant $C_y^{(2)}$ such that $D(p_{\phi|Y}, p_h^{(2)}) \leq C_y^{(2)}\eta + C_y h^2$.*

The proofs of theorem 1 and the previous inequality are given in appendix A. Numerical illustrations are given in Section 5.

4 Estimation of parameters

In the following section, we extend to incomplete data models defined by ODEs, the SAEM algorithm coupled with the H-M algorithm for the maximum likelihood estimation, and the Gibbs sampling algorithm for the Bayesian approach, to estimate the parameters $\theta = (\beta, \sigma^2)$.

4.1 Maximum likelihood approach

The EM algorithm proposed by Dempster et al. (1977) maximizes the $Q(\theta|\theta') = E(\log p_{Y,\phi}(\cdot;\theta)|y;\theta')$ function in two steps. At the k^{th} iteration, the E-step is the evaluation of $Q_k(\theta) = Q(\theta \,|\, \theta_k)$ while the M-step updates θ_k by maximizing $Q_k(\theta)$. For cases where the E-step has no closed form, Delyon et al. (1999) introduce a stochastic version SAEM of the EM algorithm. The $Q_k(\theta)$ integral is evaluated by a stochastic approximation procedure. The E-step is divided into a simulation step (S-step) of the non-observed data ϕ_k with the conditional distribution $p_{\phi|Y}(.;\theta_k)$ and a stochastic approximation step (SA-step):

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k \left(\log(p_{Y,\phi}(.;\theta_k)) - Q_k(\theta)\right),$$

where (γ_k) is a sequence of positive numbers decreasing to 0. They prove the convergence of this algorithm under general conditions in the case where $p_{Y,\phi}$ belongs to a regular curved exponential family. When the non-observed data cannot be directly simulated, Kuhn and Lavielle (2004) suggest using a MCMC scheme by building a Markov chain with $p_{\phi|Y}(\cdot;\theta_k)$ as unique stationary distribution at the k^{th} iteration.

When the regression function is defined by ODE, SAEM is implemented on the model $\mathcal{M}_h$. Let $\pi(\cdot,\beta)$ be such that $p_{h,Y,\phi}$ belongs to the exponential family:

$$p_{h,Y,\phi}(\cdot;\theta) = \exp\left\{-\Psi(\theta) + \langle S_h(y,\phi), \Phi(\theta)\rangle\right\},$$

where ψ and Φ are two functions of θ, $\langle\cdot,\cdot\rangle$ is the scalar product and $S_h(y,\phi)$ is known as the minimal sufficient statistics of the complete model, taking its value in a subset $\mathcal{S}$ of $\mathbb{R}^m$. Let $\pi(\cdot;\beta)$ be of class $\mathcal{C}^m$. At the k^{th} iteration, the SAEM algorithm is:

- S-Step: the non-observed data ϕ_k is simulated by the H-M algorithm developed in section 3 with $p_{h,\phi|Y}(\cdot;\theta_k)$ as unique stationary distribution,
- SA-Step: s_{k+1} is updated by the stochastic approximation:

$$s_{k+1} = s_k + \gamma_k(S_h(y,\phi_k) - s_k),$$

- M-Step: θ_k is updated by

$$\theta_{k+1} = \arg\max_{\theta} \left(-\Psi(\theta) + \langle s_{k+1}, \Phi(\theta)\rangle\right).$$

Kuhn and Lavielle (2004) propose estimates of the Fisher information matrix, using the Louis's missing information principle (Louis, 1982), either by importance sampling or by stochastic approximation. We adapt their estimates when the regression function is not known analytically and, as a consequence, the extended SAEM supplies the standard errors of the estimates.

The convergence of SAEM is proved on $\mathcal{M}_h$ and the distance between the likelihoods of the two models is quantified in the following theorem.

Theorem 2 *Let us consider a numerical scheme of step size h and order p. Let **H1**, **H2** and **H3** hold. Let (γ_k) be a sequence of positive numbers decreasing to 0 such that for any k in $\mathbb{N}$, $\gamma_k \in [0,1]$, $\sum_{k=1}^{\infty} \gamma_k = \infty$ and $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.*

1. *Assuming the sequence $(s_k)_{k\geq 0}$ takes its values in a compact set of $\mathcal{S}$, the sequence $(\theta_k)_{k\geq 0}$ obtained by the SAEM algorithm on $\mathcal{M}_h$, converges almost surely towards a (local) maximum of the likelihood $p_{h,Y}(y)$.*

2. *For any $\sigma_0^2 > 0$, there exists a constant θ-independent C such that*

$$\sup_{\theta=(\beta,\sigma^2)\ |\sigma^2>\sigma_0^2} |p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq Ch^p.$$

Hence, as a principal consequence of this theorem, and assuming regularity hypotheses on the Hessians of the likelihoods of both models $\mathcal{M}$ and $\mathcal{M}_h$, the bias of the estimates induced by both the numerical approximation and the estimation algorithm, is controlled.

Proof 1 *1. Assumptions of convergence of the SAEM algorithm are checked by the model $\mathcal{M}_h$. See Kuhn and Lavielle (2004) for more details.*

2. In the proof of theorem 1, we obtain the result (1b) for a fixed θ and small enough h that:

$$|p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq \frac{C_y}{(2\pi\sigma^2)^{J/2}} h^p.$$

Consequently, for any $\sigma_0^2 > 0$, for any $\theta \in \Theta$ with $\sigma^2 \geq \sigma_0^2$, there exists a constant C, independent of θ, such that

$$|p_Y(y;\theta) - p_{h,Y}(y;\theta)| \leq Ch^p.$$

4.2 Bayesian approach

The Bayesian model is defined as follows:

$$\begin{aligned} y_j &= g(t_j, \phi) + \varepsilon_j \qquad 1 \leq j \leq J \\ \varepsilon_j &\sim \mathcal{N}(0, \sigma^2) \\ \phi &\sim \pi(\cdot; \beta) \\ \theta &\sim p_\theta(\cdot; \Gamma) \end{aligned}$$

where $\theta = (\beta, \sigma^2)$ is distributed from the prior distribution p_θ with fixed hyperparameters Γ. For example, a Gamma prior distribution can be chose for σ^{-2}. The prior distribution on β depends on the specific distribution π. The Bayesian approach consists in the evaluation of the posterior distribution $p_{\theta|Y}$. The iterative Gibbs sampling algorithm is outlined as follows (see Huang et al., 2004, for more details):

- STEP 1: initialize the iteration counter of the chain $k = 1$ and start with initial values $\sigma^{-2(0)}, \beta^{(0)}, \phi^{(0)}$.
- STEP 2: obtain a new value $\sigma^{-2(k)}, \beta^{(k)}, \phi^{(k)}$ from $\sigma^{-2(k-1)}, \beta^{(k-1)}, \phi^{(k-1)}$ through successive generation of values
  1. $\sigma^{-2(k)} \sim p(\sigma^{-2}|\beta^{(k-1)}, \phi^{(k-1)}, y)$ $\beta^{(k)} \sim p(\beta|\sigma^{-2(k)}, \phi^{(k-1)}, y)$
  2. $\phi^{(k)} \sim p(\phi|\sigma^{-2(k)}, \mu^{(k)}, \Omega^{(k)}, y)$
- STEP 3: change the counter from k to $k+1$ and return to STEP 2 until convergence is reached.

For a Gamma prior distribution on σ^{-2}, the conditional distribution $p(\sigma^{-2}|\beta^{(k-1)}, \phi^{(k-1)}, y)$ is a Gamma distribution. The conditional distribution $p(\beta|\sigma^{-2(k)}, \phi^{(k-1)}, y)$ depends on the specific form of the distribution $\pi(.; \beta)$. To generate a realization of ϕ, Bennet et al. (1996) describe several approaches, such as a Rejection Gibbs, a Ratio Gibbs, independent or Random-walk H-M algorithms. Gilks et al. (1996) recommend the use of initial iterations allowing a "burn-in" phase, followed by a large number of iterations.

For models defined by ODEs, the inclusion of such H-M algorithms requires the evaluation of $g(\phi, t_j)$ at each iteration. Hence, the Gibbs algorithm is implemented on the approximate statistical model $\mathcal{M}_h$.

In practice, we estimate the posterior distribution of the model $\mathcal{M}_h$ instead of the distribution of interest $p_{\theta|Y}$. The following theorem quantifies the total variation distance between the posterior distribution $p_{h,\theta|Y}$ and this original distribution of interest, $p_{\theta|Y}$.

Theorem 3 *Let us consider a numerical scheme of step size h and order p. Let $p_{\theta|Y}$ and $p_{h,\theta|Y}$ be the posterior distributions respectively of $\mathcal{M}$ and $\mathcal{M}_h$. Assume that **H1**, **H2** and **H3** hold.*

1. *The Gibbs sampling algorithm converges on the model $\mathcal{M}_h$.*
2. *There exists a $\mathbf{y}$-dependent constant C_y such that*

$$D(p_{h,\theta|Y}, p_{\theta|Y}) \leq C_y h^p.$$

Hence, as a principal consequence of this theorem, the bias on the posterior mean is controlled: under moments hypotheses on $p_\theta(\theta)$ and $p_{h,\theta|Y}(\theta)$, there exists a constant C'_y such that $\left|E_{\theta|y}(\theta) - E_{h,\theta|y}(\theta)\right| = \left|\int \theta p_{\theta|Y}(\theta)d\theta - \int \theta p_{h,\theta|Y}(\theta)d\theta\right| \leq C'_y h^p$ where $E_{\theta|y}(\cdot)$ and $E_{h,\theta|y}(\cdot)$ are the expectation under the posterior distributions $p_{\theta|y}$ and $p_{h,\theta|y}$ respectively. Similar result can be obtained for the bias of the posterior mode, assuming regularity hypotheses on the distributions $p_{Y|\theta}$ and $p_{h,Y|\theta}$ of both models $\mathcal{M}$ and $\mathcal{M}_h$.

Proof 2 *1. Assumptions of convergence of the Gibbs sampling algorithm are checked on the model $\mathcal{M}_h$, see Carlin and Louis (2000) for more details.*

2. By Bayes theorem, we have

$$p_{\theta|Y} = \frac{p_{Y|\theta} p_\theta}{p_Y}$$

where $p_Y = \int p_{Y|\theta} p_\theta d\theta$, and the same equality for the model $\mathcal{M}_h$. From the result (1b) of the proof of theorem 1, we deduce that there exists a constant C, independent of θ, such that for any $\theta \in \Theta$ with $\sigma^2 > \sigma_0^2 > 0$, $|p_{Y|\theta}(y) - p_{h,Y|\theta}(y)| \leq Ch^p$ and $|p_Y(y) - p_{h,Y}(y)| \leq Ch^p$. We now bound $|p_{\theta|Y}(\theta) - p_{h,\theta|Y}(\theta)|$:

$$\begin{aligned} |p_{\theta|Y}(\theta) - p_{h,\theta|Y}(\theta)| &\leq \frac{p_\theta(\theta)}{|p_Y(y)|}\left||p_{Y|\theta}(y) - p_{h,Y|\theta}(y)| + \frac{p_Y(y)}{p_{h,Y}(y)}|p_Y(y) - p_{h,Y}(y)|\right| \\ &\leq \frac{Ch^p}{|p_Y(y)|} p_\theta(\theta)\left|1 + \frac{p_{h,Y|\theta}(y)}{p_{h,Y}(y)}\right| = \frac{Ch^p}{|p_Y(y)|}\left(p_\theta(\theta) + p_{h,\theta|Y}(\theta)\right). \end{aligned}$$

Thus

$$D(p_{\theta|Y}, p_{h,\theta|Y}) \leq \frac{Ch^p}{|p_Y(y)|}.$$

5 Application to nonlinear mixed-effects model

Nonlinear mixed-effects models are widely used in pharmacokinetics (PK) and pharmacodynamics (PD) to estimate PK/PD parameters. They are interesting because of their capacity to discriminate the intra- from the inter-subject variabilities and to test covariate effect on the PK/PD parameters. They are modeled by:

$$\begin{aligned} y_{ij} &= C(t_{ij}, \phi_i) + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2) \\ \phi_i &\sim \pi(.; \beta), \end{aligned}$$

where y_{ij} is the observation of the drug concentration C for subject i, $i = 1, \ldots, N$, at time t_{ij}, $j = 1, \ldots, n_i$ and ϕ_i is the vector of individual non-observed PK/PD parameters of subject i. A Gaussian distribution $\pi(.; \beta) = \mathcal{N}(\mu, \Omega)$ is classically chosen for ϕ, resulting in $\beta = (\mu, \Omega)$. In the Bayesian framework, Gamma prior distribution is chosen for σ^{-2}, Wishart or Gamma prior distribution is chosen for Ω^{-1} and a Gaussian prior distribution is used for μ.

We consider a case where the drug concentration C is defined through a differential equation without analytical solution. On a simulated dataset, we illustrate applications of the H-M, the SAEM and the Bayesian Gibbs algorithms.

5.1 Numerical settings

The following one-compartment pharmacokinetic system with a first order absorption and a Michaelis-Menten saturable elimination describes the concentration C of a drug:

$$\frac{dC}{dt}(t, \phi) = \frac{k_a D}{V} e^{-k_a t} - \frac{V_m C(t, \phi)}{k_m + C(t, \phi)}, \tag{3}$$

where D is the known administered dose, V the total volume of distribution, k_a the absorption constant, k_m the Michaelis-Menten constant and

V_m the maximum rate of metabolism. Hence, the parameter vector is $\phi = [V, k_a, k_m, V_m] \in \mathbb{R}^4$.

We consider the pharmacokinetic parameters of hydroxurea, an anticancerous drug, studied by Tracewell et al. (1995): V=12.2 L, k_a=2.72 h^{-1}, k_m=0.37 mmol/L, V_m=0.082 mmol/h/L. One dataset of 20 patients is simulated with a dose $D = 13.8$ mmol and measurements at time $t = 0, 0.5, 1, 1.5, 2$ hours and then every hour until 12 hours. We choose a Gaussian distribution for ϕ with a diagonal variance-covariance matrix (diagonal components equal to 0.4). The residual variance σ^2 is set to 0.01. The numerical method used to simulate this dataset is the ode45 solving Matlab function, that implements a Runge-Kutta scheme of the fourth order, with a very small maximal step size of resolution h equal to 0.001. Data are plotted on figure 1.

[Figure 1 about here.]

5.2 Results

Several numerical integration methods included in the H-M algorithm are compared on the simulation of the conditional distributions. 1000 independent 200-long Markov chains are generated for each method. The first method is the ode45 function with maximal step size $h = 0.001$ in order to obtain an exact cumulative distribution function; this is the reference as this numerical method is of the fourth order with a very small step size. Then we compare the two cumulative distribution functions obtained by using at first ode45 with the default step size $h = 0.1$ and then the LL and LL2 schemes. As seen on Figure 2, plotting the empirical cumulative distribution functions, the three numerical methods provide similar simulated conditional distributions.

[Figure 2 about here.]

The SAEM and Bayesian Gibbs algorithms are applied to the simulated dataset. We implemented the SAEM algorithm in Matlab and the Bayesian Gibbs algorithm is the one implemented in WinBugs and its Differential Interface software (Spiegelhalter et al., 1996). These estimates are compared with those obtained by the NONMEM software, proposed by Beal and Sheiner (1982) and used by 80% of the pharmacokineticists in drug companies. NONMEM is an implementation of the First Order Conditional Estimate algorithm, which is based on a first order linearization of the regression

function around the conditional estimates of the parameters ϕ, the standard errors of the estimates being evaluated by linearization.

Initial values have been arbitrarily chosen and are presented in Table 1. The evolutions of each SAEM parameter estimate are plotted against iterations on Figure 3. The estimates converge rapidly to a neighborhood of the simulated values.

[Figure 3 about here.]

In the Bayesian framework, a Gamma prior distribution has first been used on Ω^{-1} leading to unacceptable convergence graphs. A Wishart prior distribution has then been used on Ω^{-1} with satisfactory convergence and posterior density graphs after 100000 iterations.

The parameter estimates and their standard errors obtained by SAEM, NONMEM and the Bayesian Gibbs algorithms are presented in Table 1.

[Table 1 about here.]

In this example, it takes about ten minutes for SAEM to converge using a conventional Intel Pentium IV 3,2 GHz workstation. All the SAEM estimates almost reach the simulated values. The SAEM algorithm achieves the evaluation of the information Fisher matrix, and almost all the standard errors of the estimates are satisfactory. Using the same computer, the NONMEM software stops after about ten minutes without convergence towards the maximum of the likelihood. The estimate of V_m does not change from its initial value, the estimate of k_m is far from its simulation value, var k_m is estimated near zero while var V_m and var k_a are overestimated. The NONMEM software fails to evaluate the standard errors of all estimates. Using the same computer, it takes about one hour for the Winbugs software to compute 100 000 iterations of one Markov chain of the Gibbs algorithm. All the Bayesian estimates almost reach the simulated values and the standard errors of the estimates are satisfactory.

This simulated example illustrates the ability of the proposed algorithms to estimate precisely the parameters of mixed models defined by ODE.

6 Discussion

This paper extends the statistical approaches used to estimate incomplete data model parameters to the frequent cases where such a model $\mathcal{M}$ is defined by a dynamical system. To that purpose, an approximate model $\mathcal{M}_h$ is

introduced, of which regression function is evaluated by a numerical integration method. The standard estimation algorithms are adapted to estimate this approximate model. The convergence of the H-M, the SAEM and the Gibbs Sampling algorithms on the model $\mathcal{M}_h$ is proved.

This paper quantifies the error induced by the use of a numerical solving method. The errors on the conditional distribution, the likelihood and the posterior distribution between the model $\mathcal{M}$ and the approximate model $\mathcal{M}_h$ are controlled by h^p, where h is the step size and p the order of the numerical integration method. This error is distinct from the error on the estimates induced by the estimation algorithms, which is classically controlled by the standard errors evaluated through the Fisher information matrix of the estimates.

This paper proposes an extended version of the LL scheme, using the dependence to the parameter ϕ of the ODE, which allows to reduce significantly the computational cost of the H-M algorithm in practice. The major advantage of the LL schemes is its behavior on stiff dynamical systems, providing an interesting trade-off between computational cost and numerical stability when included in an iterative stochastic algorithm Biscay et al. (1996).

Regarding the maximum likelihood, we study the SAEM algorithm instead of the Monte-Carlo EM proposed by Wei and Tanner (1990) or Wu (2004). With an analytical regression function, the MCEM is computational intensive because of the large sample of non-observed data simulated at each iteration, while the Stochastic Approximation method requires the generation of only one realization of non-observed data at each iteration. Thus, due to computational time considerations, we only extend the SAEM algorithm to the case of regression function implicitly defined.

The SAEM algorithm is applied to a dataset simulated using a pharmacokinetic model defined by ODEs. The SAEM estimates are compared with those obtained by the standard estimation software NONMEM, the only available software providing estimates by maximum likelihood in nonlinear mixed models defined by ODEs. SAEM provides satisfying estimates and standard errors of the parameters, while NONMEM does not converge on this simulated example and fails to evaluate the standard errors. The estimation algorithm implemented in NONMEM is based on the linearization of the regression function. Despite the fact that Vonesh (1996) highlights problems of consistence and convergence of the estimates produced by such estimation methods, NONMEM is used by 80% of the pharmacokineticists in the pharmaceutical industry. The simulation results presented in this paper

point out the poor ability of this software to estimate the parameters in non-linear mixed model defined through ODEs. Thus we recommend using the SAEM algorithm to analyze such incomplete data problems defined by ODE. The SAEM algorithm is implemented by the MONOLIX group in a free Matlab function (http://mahery.math.u-psud.fr/∼lavielle/monolix/index.html). The extension of the monolix function to ODE models will soon be available on the same web-site.

Several methods have been suggested to simulate the non-observed data included in a Bayesian approach. Wakefield et al. (1994) sample the non-observed data using a ratio-of-uniform while Tierney (1994) proposes using a Hastings-Metropolis algorithm. Gilks et al. (1996) summarize these methods in their book. We use the Winbugs software and its Differential Interface software Spiegelhalter et al. (1996), which implements hybrid Gibbs and Hastings-Metropolis algorithms. On the simulated dataset, the posterior distributions are well estimated while the subject number was low.

In this paper, the methodology is illustrated on a mixed model issued from pharmacocinetics. However, there exist many others fields of applications. First, it can be applied in Functional Magnetic Resonance Imaging (fMRI), a neuroimaging technique using the vascular oxygenation contrast as an indirect measure of cerebral activity. Despite the extensive development of such techniques, the coupling mechanisms between neuronal activity and cerebral physiological changes -such as vascular changes- are still poorly understood. Recently, dynamical models such as the Balloon model have been introduced to explain these interactions (Buxton et al., 1998), leading to define statistical models defined by ODEs. Many others applications can be found in the framework of functional data analysis in the case where the underlying process can be described by a differential system.

Diffusion models described by stochastic differential equations (SDEs) are natural extensions to the corresponding deterministic models (defined by ODEs) to account for time-dependent residual errors and to handle real life variations in model parameters occurring over time. The two estimation methods proposed in this paper can be extended to this case, opening wide perspectives of application.

Acknowledgements

The authors are grateful to their advisor Professor Marc Lavielle for his constructive advice and help. The authors thank Professor France Mentré for her helpful comments. The authors would like to thank Rolando Biscay for helpful discussions about Local Linearisation schemes and Jean-Louis Foulley for his constructive help about Bayesian estimation.

A Convergence of the H-M:

1. **Proof of Theorem 1**

   (a) Following Tierney (1994), the H-M algorithm provides a uniformly ergodic Markov chain with $p_{h,\phi|Y}$ as invariant distribution, as soon as π is one of the proposal density.

   (b) By Bayes theorem, we have $p_{\phi|Y}(\phi) = \frac{p_{Y|\phi}(y|\phi)\pi(\phi)}{p_Y(y)}$, and

$$\left|p_{\phi|Y}(\phi) - p_{h,\phi|Y}(\phi)\right| = \frac{\pi(\phi)}{p_Y(y)}\left[\left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| + \frac{p_{h,Y|\phi}(y)}{p_{h,Y}(y)}\left|p_Y(y) - p_{h,Y}(y)\right|\right].$$

   We have $\left|p_{Y|\phi}(y|\phi) - p_{h,Y|\phi}(y|\phi)\right| = \left|p_{Y|\phi}(y)\right|\left|1 - \frac{p_{h,Y|\phi}(y)}{p_{Y|\phi}(y)}\right|$, and

$$p_{Y|\phi}(y) = \frac{1}{(2\pi\sigma^2)^{J/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{J}\left(y_j - g(t_j,\phi)\right)^2\right].$$

   Furthermore, by **H3**, we have:

$$\begin{aligned}\left|g_h(t,\phi) - g(t,\phi)\right| &= \left|H\circ f_h(t,\phi) - H\circ f(t,\phi)\right| \\ &\leq L_H \sup_{t,\phi}\left|f_h(t,\phi) - f(t,\phi)\right| \leq L_H C h^p.\end{aligned}$$

   By **H1** and the assumptions on the proposal densities, (t,ϕ) remains a.s. uniformly in the compact set $[t_0,T]\times K_1$. Thus there exist M and, M_h such that

$$M = \sup_{(t,\phi)\in[t_0,T]\times K_1}\left|g(t,\phi)\right|, \qquad M_h = \sup_{(t,\phi)\in[t_0,T]\times K_1}\left|g_h(t,\phi)\right|,$$

and, we can prove that $M_h \leq M + Ch^p$. Hence, there exists A_h such that

$$\left|(y_j - g_h(t_j,\phi))^2 - (y_j - g(t_j,\phi))^2\right| \leq A_h C h^p$$

and, we obtain a ϕ-independent bound:

$$\left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| \leq \frac{1}{(2\pi\sigma^2)^{J/2}}(e^{\frac{1}{2\sigma^2}JA_hCh^p}-1) \leq \frac{1}{(2\pi\sigma^2)^{J/2}}\left(e^{BA_hh^p}-1\right),$$

with $B = \frac{JC}{2\sigma^2}$. Consequently, by integrating the previous inequality,

$$|p_Y(y) - p_{h,Y}(y)| \leq \int \left|p_{Y|\phi}(y) - p_{h,Y|\phi}(y)\right| \pi(\phi)d\phi \leq \frac{1}{(2\pi\sigma^2)^{J/2}}(e^{BA_hh^p}-1).$$

Finally, using the previous inequalities, we have:

$$\int \left|p_{\phi|Y}(\phi) - p_{h,\phi|Y}(\phi)\right| d\phi \leq \frac{2}{p_Y(y)(2\pi\sigma^2)^{J/2}}(e^{BA_hh^p}-1).$$

Let $D(\cdot,\cdot)$ denote the total variation distance. If h is small enough, there exists a h-independent constant C_y such that:

$$D(p_{\phi|Y}, p_{h,\phi|Y}) \leq C_y h^p.$$

2. **Proof of the inequality of remark 2**

Let assume that the Markov Chain has a stationary distribution $p^{(2)}_{h,\phi}$. At step $r+1$, let $\phi^{(r)}$ be the current value and ϕ^c be the candidate of the Markov chain. By definition of the LL and LL2 schemes, we have

$$\left|f_h(t,\phi^c) - f_{h,\phi^{(r)}}(t,\phi^c)\right| = O(\|\phi^c - \phi^{(r)}\|_{\mathbb{R}^k}). \tag{4}$$

Moreover, by definition of the acceptance probability,

$$|\rho(\phi^{(r)},\phi^c) - \rho^{(2)}(\phi^{(r)},\phi^c)| \leq \frac{\pi(\phi^c)}{p_{h,Y|\phi}(\phi^{(r)})\pi(\phi^{(r)})}\left|p_{h,Y|\phi}(\phi^c) - p^{(2)}_{h,Y|\phi}(\phi^c)\right|$$

Let quote

$$M_{LL} = \sup_{(t,\phi)\in[t_0,T]\times K_1} |H\circ f_h(t,\phi)| \qquad \text{and} \qquad M_{LL2} = \sup_{(t,\phi)\in[t_0,T]\times K_1} |H\circ f_{h,\phi^{(r)}}(t,\phi)|.$$

These quantities exist as ϕ remains in the compact set K_1 by definition of the proposal density and, because $\phi \to f_0(\phi)$ is $\mathcal{C}^1$. According to (4), there exists M such that:

$$\begin{aligned}&\left|(y_j - H \circ f_h(t_j, \phi^c))^2 - \left(y_j - H \circ f_{h,\phi^{(r)}}(t_j, \phi^c)\right)^2\right| \\ &\leq \underbrace{\left(2\max|y_j| + M_{LL} + M_{LL2}\right) L_H M}_{\equiv A} \underbrace{\|\phi^c - \phi^{(r)}\|_{\mathbb{R}^k}}_{\leq \eta \text{ by property (2)}}.\end{aligned}$$

Consequently we have

$$\left|p_{h,Y|\phi}(\phi^c) - p^{(2)}_{h,Y|\phi}(\phi^c)\right| \leq \frac{1}{(2\pi\sigma^2)^{J/2}}\left(e^{\frac{1}{2\sigma^2}JA\eta} - 1\right).$$

As $p_{h,Y|\phi}$ is continuous in ϕ, there exists a constant c such that

$$\inf_{\phi \in K_1} p_{h,Y|\phi}(\phi) \geq \frac{c}{(2\pi\sigma^2)^{J/2}}.$$

By **H1**, and combining the previous inequalities, we obtain

$$|\rho(\phi^{(r)}, \phi^c) - \rho^{(2)}(\phi^{(r)}, \phi^c)| \leq \frac{b}{ca}\left(e^{\frac{1}{2\sigma^2}JA\eta} - 1\right) \leq B\eta, \tag{5}$$

for a η small enough. The transition kernel of the Markov chains simulated by the standard and, modified Hastings-Metropolis algorithms are quoted $\mathcal{K}$ and $\mathcal{K}^{(2)}$ respectively. Using the previous inequality, we have:

$$\begin{aligned}\left|\mathcal{K}(\phi; \{\phi^c\}) - \mathcal{K}^{(2)}(\phi; \{\phi^c\})\right| &= \left|\rho(\phi, \phi^c) - \rho^{(2)}(\phi, \phi^c)\right| q(\phi^c|\phi) \\ &\leq B\eta \; q(\phi^c|\phi)\end{aligned}$$

As a consequence, we have

$$\sup_{\phi \in K_1} D\left(\mathcal{K}(\phi; \cdot), \mathcal{K}^{(2)}(\phi, \cdot)\right) \leq B\eta$$

The result follows from the part 1 of this theorem, applied for the Local Linearization scheme, combined with the following lemma.

Lemma 2 *Let $\mathcal{K}$ and $\mathcal{K}^{(2)}$ be the respective transition kernels of two Markov chains defined on a space $\mathcal{E}$ and let p and $p^{(2)}$ be their respective stationary distributions. Assume that $\mathcal{K}$ supplies a uniformly ergodic chain and that there exists a constant C such that*

$$\sup_{\phi\in\mathcal{E}} D\left(\mathcal{K}(\phi;\cdot),\mathcal{K}^{(2)}(\phi,\cdot)\right) \leq C.$$

Then, there exists a constant α such that:

$$D\left(p,p^{(2)}\right) \leq \alpha C.$$

This lemma directly derives from the Poisson equality:

$$p\cdot f - p^{(2)}\cdot f = p^{(2)}(\mathcal{K}^{(2)}-\mathcal{K})\cdot Vf$$

where $Vf(x) = \sum_{n=0}^{\infty}(\mathcal{K}^n f(x) - p\cdot f)$ and $p\cdot f = \int f(x)p(dx)$ for any measurable function f.

***Remark* 3** *As suggested by the referees, we underline that no hypothesis is assumed on the stationary distribution. The result is based on the transition kernel properties.*

B The modified Local Linearization scheme

The LL scheme is based on a local linearization of the second member of ODE (1) with respect to time t and f. The new LL2 scheme is deduced using a Taylor expansion of the right term with respect to t, f and ϕ.

B.1 Principle

Let the equation (1) be solved by the LL scheme at a given ϕ_0. Let ϕ be in a bounded neighborhood of ϕ_0. Let the time interval $[t_0,T]$ be divided in N sub-intervals $[t_n,t_{n+1}[$, $t_n = t_0+nh$, $n=0,\ldots,N-1$, where h is the step size of the method. On each time interval $[t_n,t_{n+1}[$, the linearized equation deriving from the equation (1) at $(t,\,\phi)$ with $t\in[t_n,t_{n+1}[$ is:

$$\begin{aligned} F(f(t,\phi),t,\phi) \simeq\ & F(f(t_n,\phi_0),t_n,\phi_0) + \frac{dF}{df}\left(f(t_n,\phi_0),t_n,\phi_0\right)\left(f(t,\phi)-f(t_n,\phi_0)\right) \\ +\ & \frac{dF}{dt}\left(f(t_n,\phi_0),t_n,\phi_0\right)(t-t_n) + \frac{dF}{d\phi}\left(f(t_n,\phi_0),t_n,\phi_0\right)(\phi-\phi_0). \end{aligned} \tag{6}$$

The following notations are introduced:

$$\left\{\begin{array}{lcllcl} f_n(\phi) & = & f_h(t_n,\phi) & F'_n(\phi) & = & \frac{dF}{dt}\left(f_n(\phi),t_n,\phi\right) \\ F_n(\phi) & = & F(f_n(\phi),t_n,\phi) & R_k(X,h) & = & \int_0^h \exp(uX)u^k du \\ DF_n(\phi) & = & \frac{dF}{df}\left(f_n(\phi),t_n,\phi\right) \quad , & D_\phi F(\phi_0) & = & \frac{dF}{d\phi}\left(f_n(\phi_0),t_n,\phi\right). \end{array}\right.$$

The LL2 scheme is:

$$f_{n+1}(\phi) = f_n(\phi_0) + \Lambda_n^m(\phi_0,\phi,h) \tag{7}$$

with

$$\begin{aligned}\Lambda_n^m(\phi_0,\phi,h) &= \left[hR_0(DF_n(\phi_0),h) - R_1(DF_n(\phi_0),h)\right]F'_n(\phi_0) \\ &+R_0(DF_n(\phi_0),h)\left(F_n(\phi_0) + D_\phi F(\phi_0)(\phi-\phi_0)\right) + \exp(hDF_n(\phi_0))\left(f_n(\phi) - f_n(\phi_0)\right).\end{aligned}$$

***Remark* 4** *1. The previous recursive formula taken at $\phi = \phi_0$ is the same as the one deriving from the LL scheme.*

2. For autonomous dynamic system, the scheme (7) is simplified by:

$$f_{n+1}(\phi) = f_n(\phi_0) + \lambda_n^m(\phi_0,\phi,h),$$

with

$$\begin{aligned}\lambda_n^m(\phi_0,\phi,h) &= R_0(DF_n(\phi_0),h)\left(F_n(\phi_0) + D_\phi F(\phi_0)(\phi-\phi_0)\right) \\ &\quad + \exp(hDF_n(\phi_0))\left(f_n(\phi) - f_n(\phi_0)\right).\end{aligned}$$

B.2 Convergence of the method

Biscay et al. (1996) and Ramos and García-López (1997) prove that the LL scheme is of convergence rate h^2. We extend their results to the LL2 scheme:

Lemma 3 (Error estimation) *Let f be the exact solution of ODE (1) and f_{h,ϕ_0} be the appoximate solution obtained by the LL2 scheme with the step size h, given a point ϕ_0. Under assumptions **H1′** and **H2**, there exist two constants C_1 and C_2, ϕ and ϕ_0- independent such that for any $t \in [t_0,T]$ and for any ϕ,*

$$|f(t,\phi) - f_{h,\phi_0}(t,\phi)| \leq \max(C_1h^2, C_2\left\|\phi-\phi_0\right\|^2_{\mathbb{R}^k}).$$

Proof 3 *The proof, essentially the same as the Ramos and García-López (1997)'s one, is presented for $d = 1$ but is easily generalizable for any d.*

On the interval $[t_n, t_{n+1}[$, $0 \leq n \leq N-1$, f_{h,ϕ_0} is the exact solution of the following linear equation:

$$\left\{ \begin{array}{lcl} \frac{\partial f_{h,\phi_0}(t,\phi)}{\partial t} & = & F_{h,\phi_0}(f_{h,\phi_0}(t,\phi),t,\phi) \\ f_{h,\phi_0}(t_0,\phi) & = & f_0(\phi), \end{array} \right.$$

where F_{h,ϕ_0} is the Taylor expansion of F at point (t_n, ϕ_0) defined by the equation (6). For any t in $[t_n, t_{n+1}[$, we have:

$$|f(t,\phi) - f_{h,\phi_0}(t,\phi)| = \left| \int_{t_n}^{t} F(f(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi)du + f(t_n,\phi) - f_{h,\phi_0}(t_n,\phi) \right|$$

*Assuming **H2**, F is $\mathcal{C}^2$ on its domain of definition. Since F_{h,ϕ_0} is the second order Taylor expansion of F at (t_n, ϕ_0), there exists $\xi \in \mathbb{R}^d \times (t_n, t_{n+1}] \times \mathbb{R}^k$ such that:*

$$\begin{aligned} |F(f_{h,\phi_0}(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi)| &\leq \frac{1}{2}\left|\frac{\partial^2 F}{\partial f^2}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))^2\right| \\ +\frac{1}{2}\left|\frac{\partial^2 F}{\partial t^2}(\xi,\phi_0)(u-t_n)^2\right| &+ \frac{1}{2}\left|\frac{\partial^2 F}{\partial f \partial t}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))(u-t_n)\right| \\ +\frac{1}{2}\left|\sum_{m=1}^{k}\frac{\partial^2 F}{\partial f \partial \phi_m}(\xi,\phi_0)(f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0))(\phi_m - \phi_{0_m})\right| &+ \frac{1}{2}\left|\sum_{m=1}^{k}\frac{\partial^2 F}{\partial t \partial \phi_m}(\xi,\phi_0)(u-t_n)(\phi_m - \phi_{0_m})\right| \\ +\frac{1}{2}\left|\sum_{m',m=1}^{k}\frac{\partial^2 F}{\partial \phi_m \partial \phi_{m'}}(\xi,\phi_0)(\phi_m - \phi_{0_m})(\phi_{m'} - \phi_{0_{m'}})\right|. & \end{aligned} \quad (8)$$

*Assuming **H1'** and **H2**, there exists a constant c independent of t and ϕ, which upper-bounds every second order differential of F. Moreover, a short recursive argument together with **H1'** implies that f_{h,ϕ_0} is $\mathcal{C}^1$ on $[t_0, T] \times K_1$. Thus, there exist constants η and η' such that*

$$|f_{h,\phi_0}(u,\phi) - f_{h,\phi_0}(t_n,\phi_0)| \leq \max(\eta|u-t_n|, \eta'\|\phi - \phi_0\|_{\mathbb{R}^k})$$

Hence, using (8), there exist two constants A and A' such that:

$$|F(f_{h,\phi_0}(u,\phi),u,\phi) - F_{h,\phi_0}(f_{h,\phi_0}(u,\phi),u,\phi)| \leq \max(A|u-t_n|, A'\|\phi-\phi_0\|_{\mathbb{R}^k})^2.$$

The F function is $\mathcal{C}^2$. Thus, by quoting L_F its Lipschitz constant, we can write:

$$|F(f(u,\phi),u,\phi) - F(f_{h,\phi_0}(u,\phi),u,\phi)| \leq L_F \|f(u,\phi) - f_{h,\phi_0}(u,\phi)\|$$

Finally, by quoting $E_\phi(u) = \|f(u,\phi) - f_{h,\phi_0}(u,\phi)\|$, and combining the previous inequalities, we have:

$$E_\phi(t) \leq \int_{t_n}^{t} L_F E_\phi(u)du + \int_{t_n}^{t} \max(A|u-t_n|, A'\|\phi-\phi_0\|_{\mathbb{R}^k})^2 du + E_\phi(t_n).$$

As $|u-t_n| \leq| t-t_n| \leq h$, we have:

$$E_\phi(t) \leq \int_{t_n}^{t} L_F E_\phi(u)du + \max(Ah, A'\|\phi-\phi_0\|_{\mathbb{R}^k})^2 (t-t_n) + E_\phi(t_n).$$

The expected result derives from the following lemma and similar arguments as those presented by Ramos and García-López (1997):

Lemma 4 *Let u be a positive function such that*

$$u(t) \leq a + b(t-t_0) + c\int_{t_0}^{t} u(s)ds$$

Then we have

$$u(t) \leq ae^{c(t-t_0)} + \frac{b}{c}(e^{c(t-t_0)} - 1).$$

References

Beal, S., Sheiner, L., 1982. Estimating population kinetics. Crit Rev Biomed Eng 8(3), 195–222.

Bennet, J. E., Racine-Poon, A., Wakefield, J. C., 1996. MCMC for nonlinear hierarchical models, Chapman & Hall, London. 339–358.

Biscay, R., Jimenez, J. C., Riera, J. J., Valdes, P. A., 1996. Local linearization method for the numerical solution of stochastic differential equations. Ann. Inst. Statist. Math. 48(4), 631–644.

Buxton, R. B., Wong, E. C., Franck, L. R., 1998. Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. MRM 39, 855–864.

Carlin, B. P., Louis, T. A., 2000. Bayes and empirical Bayes methods for data analysis, vol. 69 of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Celeux, G., Diebolt, J., 1985. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational. Statistics Quaterly 2, 73–82.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Ann. Statist. 27, 94–128.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39(1), 1–38. With discussion.

Gelfand, A. E., Smith, A. F. M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85(410), 398–409.

Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1996. Markov chain Monte Carlo in practice. Interdisciplinary Statistics. Chapman & Hall, London.

Hairer, E., Nørsett, S., Wanner, G., 1987. Solving Ordinary Differential Equations I, Nonstiff Problems. Springer, Berlin.

Huang, Y., Liu, D., Wu, H., 2004. Hierarchical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. Technical Report 04/06 .

Jimenez, J. C., 2002. A simple algebraic expression to evaluate the local linearization schemes for stochastic differential equations. Appl. Math. Lett. 15(6), 775–780.

Jimenez, J. C., Biscay, R., Mora, C., Rodriguez, L. M., 2002. Dynamic properties of the local linearization method for initial-value problems. Appl. Math. Comput. 126(1), 63–81.

Kuhn, E., Lavielle, M., 2004. Coupling a stochastic approximation version of EM with a MCMC procedure. ESAIM P&S , 115–131.

Louis, T. A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44(2), 226–233.

Ramos, J. I., García-López, C. M., 1997. Piecewise-linearized methods for initial-value problems. Appl. Math. Comput. 82(2-3), 273–302.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., 1996. Bugs: Bayesian inference using gibbs sampling, version 0.5. Tech. rep.

Tierney, L., 1994. Markov chains for exploring posterior distributions. Ann. Statist. 22(4), 1701–1762. With discussion and a rejoinder by the author.

Tracewell, W., Trump, D., Vaughan, W., Smith, D., Gwilt, P., 1995. Population pharmacokinetics of hydroxyurea in cancer patients. Cancer Chemother. Pharmacol. 35(5), 417–22.

Vonesh, E. F., 1996. A note on the use of Laplace's approximation for nonlinear mixed-effects models. Biometrika 83(2), 447–452.

Wakefield, J., Smith, A., Racine-Poon, A., Gelfand, A., 1994. Bayesian analysis of linear and non-linear population models using the gibbs sampler. Appl. Statist 43, 201–221.

Wei, G. C. G., Tanner, M. A., 1990. Calculating the content and boundary of the highest posterior density region via data augmentation. Biometrika 77(3), 649–652.

Wu, L., 2004. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. J. Amer. Statist. Assoc. 99(467), 700–709.

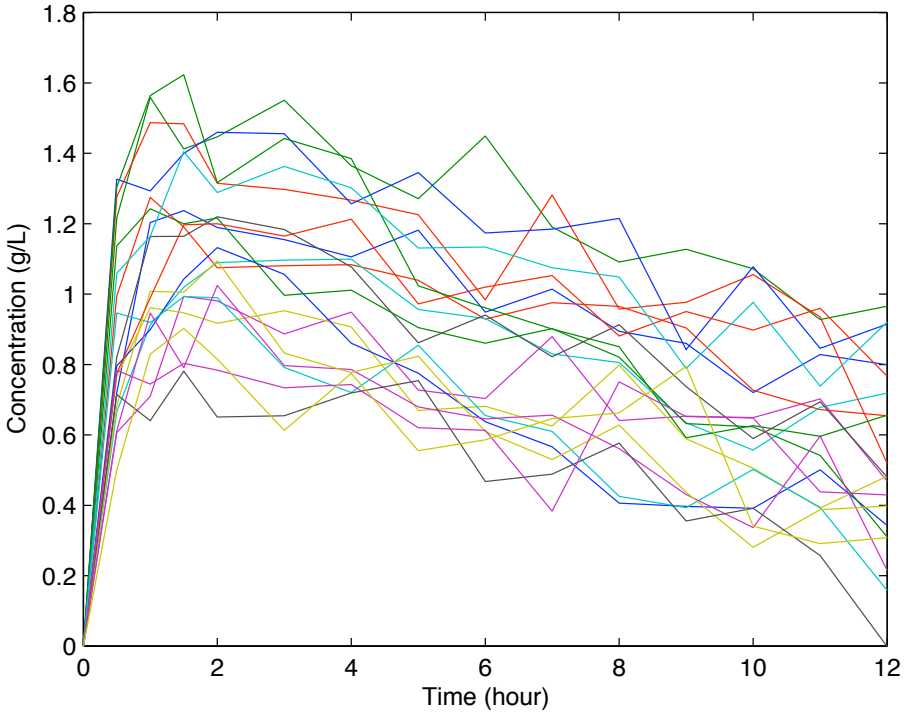

Figure 1: Individual concentrations of pharmacokinetic hydroxurea simulated for 20 patients.

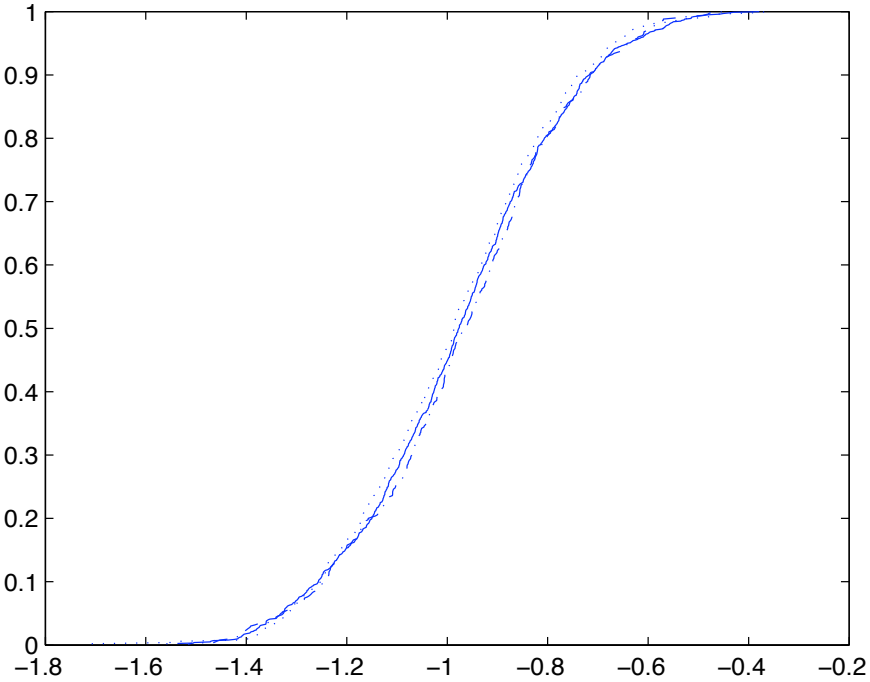

Figure 2: Empirical cumulative distribution functions of the k_m conditional distribution simulated by Hastings-Metropolis using a very precise Runge-Kutta solving scheme (plain line), a classical Runge-Kutta scheme (dotted line), or the LL and LL2 schemes (dashed line).

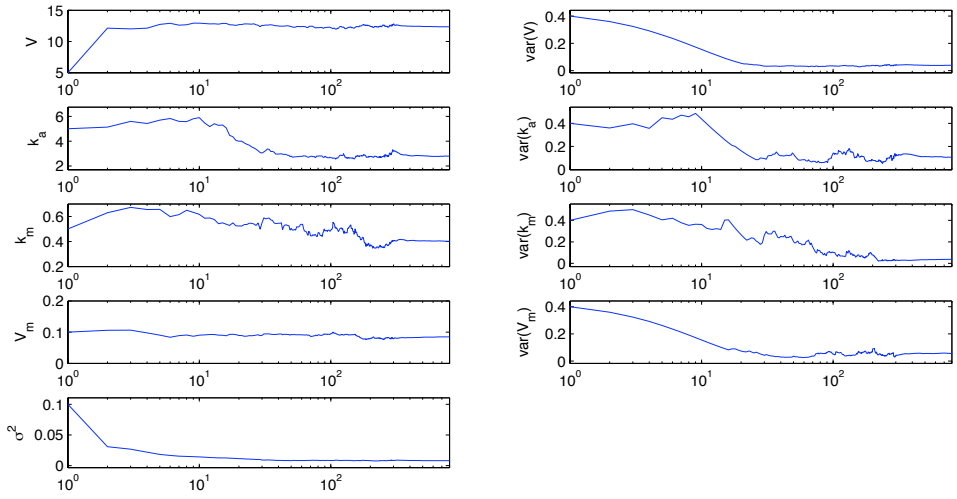

Figure 3: Evolution of the estimates in function of the iteration of SAEM algorithm (with a logarithm scale for the abscis axis).

Table 1: Parameter estimates obtained by the SAEM, NONMEM and Bayesian Gibbs algorithms on the simulated dataset.

	V	k_a	k_m	V_m	var V	var k_a	var k_m	var V_m	σ^2
initial values	5.0	5.00	0.50	0.100	0.400	0.400	0.400	0.400	0.1000
simulation values	12.2	2.72	0.37	0.082	0.040	0.040	0.040	0.040	0.0100
SAEM									
estimates	12.3	2.79	0.40	0.085	0.038	0.049	0.039	0.038	0.0081
SE	0.5	0.21	0.01	0.004	0.012	0.019	0.013	0.013	6.10^{-4}
NONMEM									
estimates	12.3	2.57	0.60	0.100	0.036	0.068	10^{-8}	0.062	0.0088
SE	-	-	-	-	-	-	-	-	-
Bayesian Gibbs									
mode estimates	12.4	2.64	0.39	0.085	0.040	0.040	0.048	0.052	0.0082
SE	0.6	0.23	0.62	0.044	0.014	0.046	0.028	0.029	7.58^{-4}